BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1583–1590
doi:10.1093/bioinformatics/bth130

An object model and database for functional genomics

Andrew Jones[1,], Ela Hunt[1], Jonathan M. Wastling[2], Angel Pizarro[3] and Christian J. Stoeckert Jr[3]*

[1]Department of Computing Science and [2]Institute of Biomedical and Life Sciences, University of Glasgow, 17 Lilybank Gardens, Glasgow, G12 8QQ, UK and [3]Center for Bioinformatics, University of Pennsylvania, 14th Floor Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA

Received on October 1, 2003; revised and accepted on December 17, 2003
Advance Access publication May 14, 2004

ABSTRACT

Motivation: Large-scale functional genomics analysis is now feasible and presents significant challenges in data analysis, storage and querying. Data standards are required to enable the development of public data repositories and to improve data sharing. There is an established data format for microarrays (microarray gene expression markup language, MAGE-ML) and a draft standard for proteomics (PEDRo). We believe that all types of functional genomics experiments should be annotated in a consistent manner, and we hope to open up new ways of comparing multiple datasets used in functional genomics.

Results: We have created a functional genomics experiment object model (FGE-OM), developed from the microarray model, MAGE-OM and two models for proteomics, PEDRo and our own model (Gla-PSI—Glasgow Proposal for the Proteomics Standards Initiative). FGE-OM comprises three namespaces representing (i) the parts of the model common to all functional genomics experiments; (ii) microarray-specific components; and (iii) proteomics-specific components. We believe that FGE-OM should initiate discussion about the contents and structure of the next version of MAGE and the future of proteomics standards. A prototype database called RNA And Protein Abundance Database (RAPAD), based on FGE-OM, has been implemented and populated with data from microbial pathogenesis.

Availability: FGE-OM and the RAPAD schema are available from http://www.gusdb.org/fge.html, along with a set of more detailed diagrams. RAPAD can be accessed by registration at the site.

Contact: jonesa@dcs.gla.ac.uk

INTRODUCTION

Proteomics uses experimental techniques for the large-scale study of proteins. The experiments aim to determine the expression of all proteins in a particular sample, search for protein–protein interactions, or use immunohistochemistry to localize the position of expression. Proteomics is a part of functional genomics, which includes microarray analysis, phenotypic studies and small molecule arrays. The integration of all the diverse types of data is vital, and new bioinformatics tools are required (Tyers and Mann, 2003). In this work, we focus on the development of an object model to represent microarray and proteomics data, including separation techniques such as two-dimensional gel electrophoresis (2-DE) and protein identification by mass spectrometry (MS). The model also stores experimental protocols, raw data and data analysis and is known as Functional Genomics Experiment Object Model (FGE-OM). FGE-OM was developed from three main sources: the MAGE model for microarrays (Spellman *et al*., 2002), the PEDRo model developed at the University of Manchester (Taylor *et al*., 2003), and a model developed at the University of Glasgow, referred to as Gla-PSI (Jones *et al*., 2003). PEDRo is a proposal for a standard format for proteomics, covering 2-DE and MS, with limited support for describing the origin of a biological sample. Gla-PSI is a response to PEDRo, including information regarding image analysis of 2-DE, difference gel electrophoresis (Ünlü *et al*., 1997) and analysis of multiple gels.

An integrated data format will facilitate the development of public repositories for storage of and querying functional genomics data. Microarray experiments are used widely because a large number of assays can be performed concurrently, and it is believed that changes in gene expression may be indicative of changes at the protein level and hence could be functionally significant (Futcher *et al*., 1999). Proteomics experiments determine the relative level of protein produced and therefore would be expected to be a better indicator of the level of protein activity. Proteomics experiments can also give information about post-translational modifications, which may have important effects on the function of the protein. It is therefore desirable that microarray and proteomics data can be queried

*To whom correspondence should be addressed.

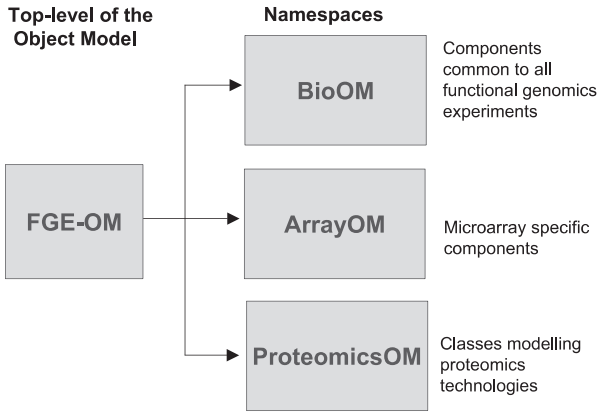

Fig. 1. An overview of FGE-OM. The model is divided into three namespaces: BioOM, ArrayOM and ProteomicsOM.

in parallel to determine the extent of gene expression and the level of encoded protein observed for a particular gene. Protein and mRNA expression data should also be accessible with genomic data to allow better annotation of the genome with functional information derived from the studies, such as protein X is up-regulated under condition Y. Databases should also incorporate information from immunohistochemistry and protein interaction studies, such as yeast two-hybrid (Fields and Song, 1989). Such systems would enable data mining applications to be developed that search for the factors that affect regulation of transcription and translation and, ultimately, protein function. Integrated databases will aid in the development of mathematical models capturing the effects of changes at the system level and could provide source data for the modelling of metabolic pathways (Voit, 2002). Data mining algorithms could then be employed to search for genes that may be important in a condition of interest, such as drug targets for a particular disease.

FGE-OM comprises three namespaces that organize the classes in logical subsets: BioOM, ArrayOM and ProteomicsOM (Fig. 1). Substantial detail from MAGE-OM has been used to develop BioOM and ArrayOM. BioOM contains a set of objects that describe protocols, sample tracking and an experimental overview from microarrays, proteomics or potentially other functional genomics techniques. ArrayOM and ProteomicsOM capture information specific to that technology. We have developed a prototype database system based on FGE-OM, known as RAPAD. RAPAD uses tables from the RAD schema for microarrays (Stoeckert *et al*., 2001) to store data from BioOM and ArrayOM. Additional tables were added to the schema derived from PEDRo and Gla-PSI to store proteomics data. RAD is a core component of Genomics Unified Schema (GUS), developed at the University of Pennsylvania. GUS is a data warehouse that incorporates genome information, expressed sequence tags (ESTs), RNA, SAGE (Velculescu *et al*., 1995), microarrays and organism specific data (http://www.gusdb.org/). The RAPAD schema has been deployed and tested for expressivity and query performance using data from microbial pathogenesis. In the future, a proteome-specific namespace will be added to GUS alongside genomic and transcriptomic components, as a step towards achieving a fully integrated database for functional genomics.

FGE-OM will be converted into an XML schema (http://www.w3.org/XML/Schema) to enable data produced by different research groups to be validated. A programming interface is already available for formatting microarray data into MAGE-ML (http://www.mged.org/Workgroups/MAGE/magestk.html), an XML implementation of MAGE-OM. MAGE-ML acts as a format for sending data to publicly available databases such as ArrayExpress (Brazma *et al*., 2003) and RAD. We are developing software capable of formatting microarray and proteomics data into FGE-ML, an XML implementation of FGE-OM. The definition of an XML Schema will ensure that data produced by different research groups are formatted in a consistent manner, can be exchanged more easily and sent to centralized databases for publication.

SYSTEMS AND METHODS

FGE-OM is expressed in Unified Modeling Language (UML; http://www.uml.org/), which is a standard notation designed to improve the process of developing large software systems (Rumbaugh *et al*., 1999). UML includes notation to represent the design and visualization of the architecture of a system during development. UML supports the definition of use case scenarios and workflows that can model the biological research process and can also be used for database design. We use class diagrams of UML to represent the concepts, objects and relationships in microarray and proteomics experiments. Class diagrams represent real world objects as a set of classes with attributes of certain types (such as strings, integers, or user defined), and relationships between classes (Fig. 2). An object model enables developers to have a shared understanding of the components of a complex system but can also be converted into an XML representation and a database implementation without significant effort.

The use of controlled vocabularies and ontologies is essential for unambiguous representation of functional genomics experiments. For MAGE-OM, it was recognized that these may come from different sources but that their usage should be explicitly indicated. This was accomplished by the creation of the *OntologyEntry* class in MAGE-OM and retained in FGE-OM (e.g. Fig. 9). The MGED Ontology (MO; Stoeckert and Parkinson, 2003) covers all the current needs of MAGE and will be used for the BioOM and ArrayOM namespaces. We have used MO to populate data entry forms for RAPAD. We will develop an ontology, in collaboration with the PSI, to describe terms used in protein expression studies, for example the types of protein modification observed. One of

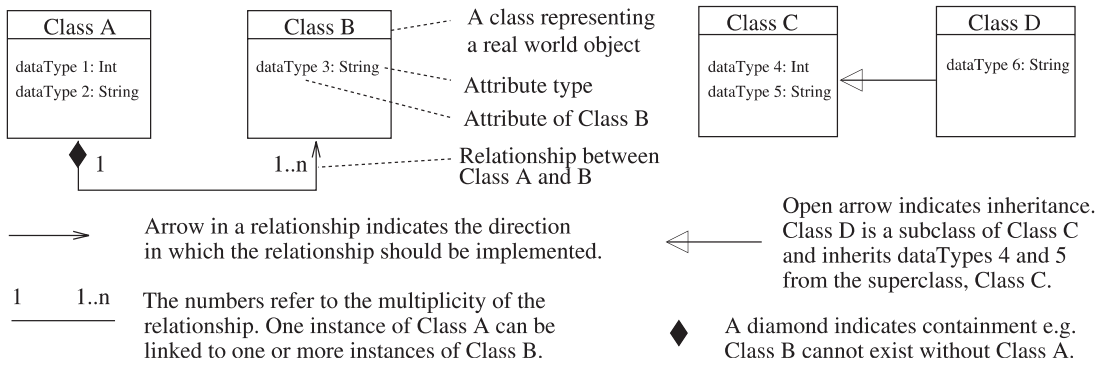

Fig. 2. The main components of a UML class diagram.

the strengths of the MAGE-OM notion of *OntologyEntry* is that the usage of controlled vocabularies is not 'hard-coded' but the source can be explicitly specified. This allows taking advantage of ongoing standards efforts in other fields.

Overview of FGE-OM

FGE-OM models microarray and proteomics data. We start with the description of a sample biological workflow, as observed in the laboratories we work in. The components of the three namespaces in FGE-OM are also described: BioOM, ArrayOM and ProteomicsOM. A complete listing of the classes in the three namespaces, along with more detailed diagrams can be found on our Web site.

A workflow for proteomics

A sample workflow is displayed in Figure 3, demonstrating how FGE-OM captures proteomics data. The overview of the experiment is modelled by the class *Experiment*. If the experiment includes multiple samples, e.g. comparing a number of 2D gels, the parameter that is varied between samples is attached to classes referencing *Experiment*. A biological substance must be processed to extract proteins and make the proteins soluble in a multi-stage process. This is modelled by a series of treatments (*Treatment*) applied to a substance (*BioMaterial*), to produce the final soluble mixture of proteins, on which certain separation techniques may be performed. Separation techniques such as 2-DE or liquid chromatography are modelled as specialized subclasses of *BioAssayTreatment*. Each *BioAssayTreatment* has a measured source of material, which is stored in *BioMaterial* and *BioMaterialMeasurement*. When data are produced by imaging a 2D gel, an instance of *PhysicalBioAssay* is created. *PhysicalBioAssay* can be referenced by the class *ImageAcquisition*, representing the scanning of the gel. 2-DE image analysis is represented by *GelImageAnalysis*, which is a subclass of *FeatureExtraction*. Gel spot data produced by image analysis can be stored in specific classes in ProteomicsOM, linked to image acquisition via *MeasuredBioAssay*.

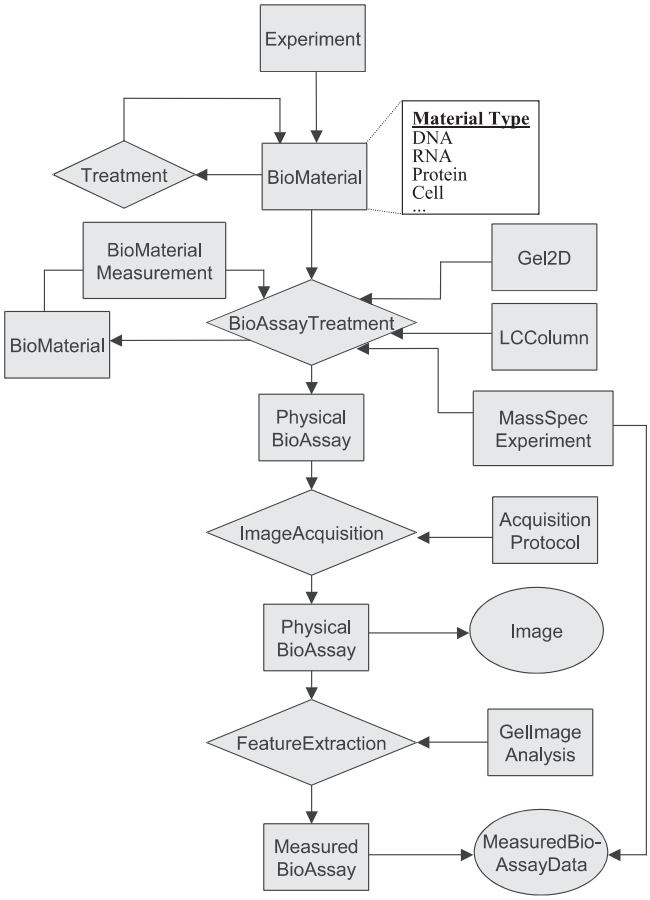

Fig. 3. A workflow for a proteomics experiment involving 2-DE or liquid chromatography to separate proteins, followed by MS to identify proteins. Diamonds indicate events, rectangles are physical entities and ovals represent data.

If MS is performed on a spot excised from a gel, or a fraction from a column, an instance of *BioMaterial* is created. *MassSpecExperiment* is a subclass of *BioAssayTreatment*, which can be linked to the source of material. MS data obtained from a particular gel spot are linked directly to

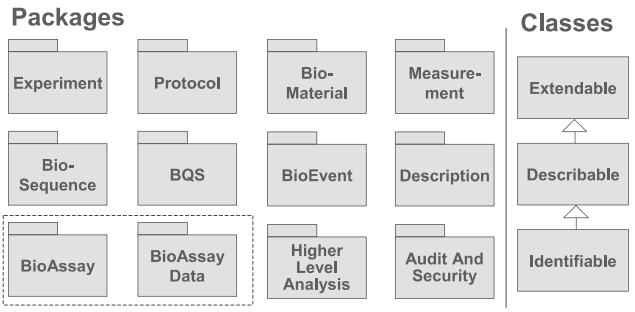

Fig. 4. The packages and classes in the BioOM namespace of FGE-OM. The boxed packages have been altered from MAGE-OM; others are identical to packages in MAGE-OM.

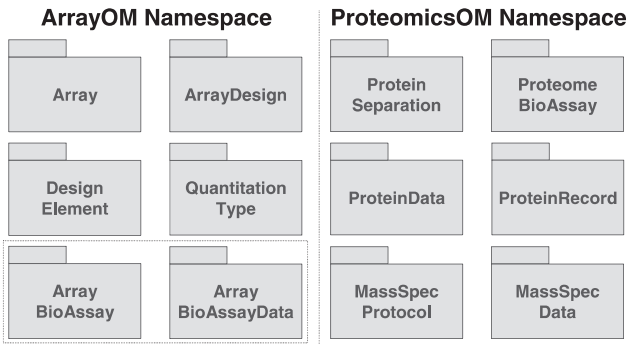

Fig. 5. The packages in the ArrayOM and ProteomicsOM namespaces. The boxed packages are newly created in FGE-OM but contain a number of classes derived from MAGE-OM. The other packages in ArrayOM are identical to packages with the same name in MAGE-OM.

data produced by image analysis of the spot, captured in *MeasuredBioAssayData*.

BioOM

Figure 4 shows the packages that constitute the BioOM namespace, which covers the components in FGE-OM that are common to all experiment types. The majority of the packages are identical to packages of the same name in MAGE-OM. However, components of the packages BioAssay and BioAssayData, which contain array-specific information, have been placed in newly created packages within the ArrayOM namespace. The three abstract classes at the top-level: *Extendable*, *Describable* and *Identifiable* are unchanged from MAGE-OM, and most classes inherit their attributes. *Identifiable* allows a name and an identifier to be added to classes. *Describable* enables links to external ontologies, data ownership and an audit trail to be attached. *Extendable* enables a triplet of attributes, Name, Value, Type, to be attached to any class for storage of properties that cannot be covered in other parts of the model.

The BioAssay package in MAGE-OM contains a class describing the hybridization of mRNA to an array. We propose to move this class to ArrayOM, and have created a new package in ArrayOM containing the *Hybridization* class. We have named this package ArrayBioAssay. The rest of the classes in BioOM:BioAssay are the same as in MAGE-OM. The BioOM:BioAssayData package contains only five classes: *BioAssayData*, *BioAssayDimension*, *MeasuredBioAssayData*, *BioDataTuples* and *BioDataValues*. The five classes are identical to those in MAGE-OM. These classes specify the general structure and location of data from any type of experiment and therefore reside in the BioOM namespace. *BioAssayDimension* allows experimental data to be packaged together across a range of conditions such as multiple array or multiple gel comparison. Classes containing information specific to microarray data reside in a newly created package, ArrayBioAssayData in ArrayOM, and are linked to classes in BioOM:BioAssayData. Technology specific details of proteome data are stored in the package ProteinData in ProteomicsOM.

ArrayOM

The ArrayOM namespace (Fig. 5) contains the packages derived from MAGE-OM that are microarray specific. The packages Array, ArrayDesign and DesignElement describe the layout of features on a microarray and have not been altered. QuantitationType includes details of how array data are quantified and is therefore also included in ArrayOM. However, various data types from functional genomics experiments could be quantified in similar ways, using standard statistical tests. Therefore, in the future, a new package could be added to BioOM to model statistical processing, recording the software used and the parameters employed. The ArrayBioAssay package contains only *Hybridization*, which is linked to classes in BioOM:BioAssay. The ArrayBioAssayData package is a modified version of the BioAssayData package in MAGE-OM. ArrayBioAssayData includes the MAGE-OM derived class *BioDataCube* which, represents the three dimensions of data: the array features; the parameter that is varied across a multiple array experiment; and the values calculated for each array feature, such as the relative fluorescence. *BioDataCube* captures the order of the three dimensions and stores pointers to separate files containing large quantities of numerical data. The three dimensions of data also exist in a proteomics experiment, and potentially in other functional genomics experiments, and therefore in theory it should be possible to create a generic data model in BioOM that models the dimensions of data. However, we believe that *BioDataCube* is possibly too simplistic to capture proteomics data, having only an ordering and pointers to lists of values in files. In proteomics, a multiple 2-DE experiment may detect certain proteins present on one gel and not another, calculated by image analysis software. The comparison of multiple gels can be error prone, and spots matched

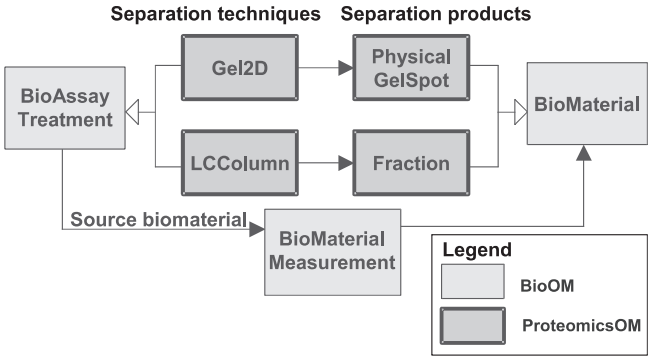

Fig. 6. The ProteinSeparation package contains classes that model the relationship between separation techniques and products.

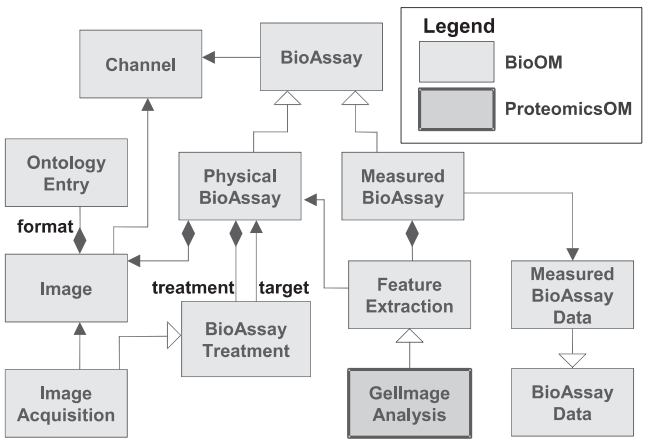

Fig. 7. The relationship between *GelImageAnalysis*, in the ProteomeBioAssay package, with classes from the BioAssay package in BioOM.

across multiple gels may have scores assigned to the quality of the match. Spots may also be matched based on experimental evidence, such as MS data. We therefore believe that a generic data model covering all types of functional genomics experiments would have to be more complex and are planning to develop one in the future.

ProteomicsOM

The proteomics namespace (Fig. 5) is a further development of the PEDRo and Gla-PSI models. The design of PEDRo was based upon principles different from the design of MAGE-OM. MAGE-OM was intended to be future proof, by including generic attributes in classes and allowing data types to be specified using controlled vocabularies of terms, rather than specifying explicitly in the model which data types should be stored in which position. PEDRo contains specific named attributes for all the data types that may need to be recorded. For example, in 2-DE, a gel is used to separate thousands of proteins into individual spots. An image of the gel is analysed with specialized software that produces output about gel spots, such as an estimate of volume, area, the coordinates on the gel and many others. PEDRo aims to define explicitly all the data types that are created by image analysis software. A model following MAGE design principles would have a placeholder for the first data type and value, followed by the second data type and value and so on. ProteomicsOM includes the classes from PEDRo in new packages, however, the classes have been linked to components in BioOM that allow generic protocols and parameters to be attached, as required. The following sections describe the classes that are contained within the six packages of ProteomicsOM.

ProteinSeparation package

The ProteinSeparation package describes a number of separation techniques, including 2-DE and liquid chromatography (summarized in Fig. 6). Classes modelling separation techniques are subclasses of *BioAssayTreatment* within BioOM. An instance of *BioAssayTreatment* can be linked to *Protocol*, which allows any type of protocol information to be added regarding hardware or software, along with a set of parameters. Therefore, classes derived from PEDRo can acquire new parameters if the attributes specified in the model do not cover the information that must be stored. This mechanism will be particularly important for storing information about new technologies that cannot be covered by PEDRo as it stands. The products of a separation technique, such as a gel spot or column fraction, are modelled as classes, with attributes derived from PEDRo, and are subclasses of *BioMaterial*. A separation product can become the input for another separation technique, and therefore the model utilizes a link from *BioAssayTreatment* to *BioMaterial* via *BioMaterialMeasurement* to specify the source of material. These three classes are all contained within BioOM.

ProteomeBioAssay package

The ProteomeBioAssay package contains only one class, *GelImageAnalysis*; however, new relationships have been added to enable the reuse of classes in BioOM:BioAssay in the protein context (Fig. 7). These relationships have the following semantics. *FeatureExtraction* from MAGE-OM models the process by which data are extracted from a scanned microarray. In ProteomicsOM, *GelImageAnalysis* is a subclass of *FeatureExtraction* and models the process of analysing a 2D gel with specialist software. *FeatureExtraction* is linked to *PhysicalBioAssay*, which is linked to the source image (*Image*), the scanning process (*ImageAcquisition*) and information about a specific channel or wavelength at which the array has been scanned (*Channel*). These classes can be re-used in proteomics to refer to the scanning of a 2D gel. We reuse *Channel* to model the technique of difference gel electrophoresis, in which a single gel is scanned at a number of different wavelengths. Data that are obtained from image

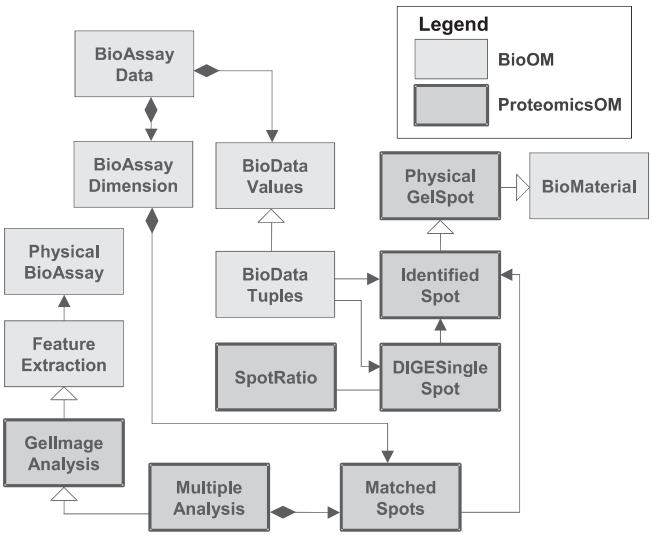

Fig. 8. The ProteinData package.

analysis are stored in classes linked to *BioAssayData* in the ProteinData package.

ProteinData package

The ProteinData package models information about gel spots (Fig. 8). Spot data are captured in *IdentifiedSpot*, with attributes covering data types produced by image analysis software. The model also captures data from difference gel electrophoresis: single channel spots in *DIGESingleSpot*, and co-migrated spots (from the composite image) are stored in *IdentifiedSpot*. Spot data are linked to the gel from which they were produced, as *IdentifiedSpot* is a subclass of *PhysicalGelSpot*, which is directly linked to Gel2D in the ProteinSeparation package. Spot data are linked back to the image analysis from which they were produced via *BioAssayData* and *MeasuredBioAssay* (Fig. 7). The ProteinData package also captures multiple gel comparisons. *BioAssayDimension* in BioOM models multiple sample comparisons and is used in ProteomicsOM by the addition of a link to *MatchedSpots*, modelling spots matched across multiple gels to capture differential expression of proteins.

MassSpecProtocol and MassSpecData packages

The packages capturing MS data and protocols contain classes derived from PEDRo (Fig. 9). MS protocols are modelled by a package called MassSpecProtocol, which contains a class at the top level called *MassSpecExperiment*. *MassSpecExperiment* is a subclass of BioAssayTreatment, which can be used to link to the biological substance on which MS has been performed (in *BioMaterial*). The substance can be the product of a series of separation techniques. PEDRo-derived classes specify many of the parameters that are associated with an MS instrument, along with details of the MS ion source. Additional text and parameters not covered in these classes can be

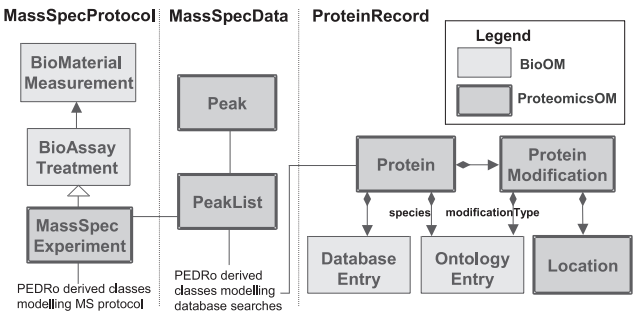

Fig. 9. The model of MS data and protocols linked to the ProteinRecord package.

attached using the generic *Protocol* class in BioOM, linked to *BioAssayTreatment*. This ensures that the model can be extended to include protocols from different MS instrument manufacturers, new software and new technologies. We define a new package, MassSpecData, to store lists of peaks from a trace and database searches. Proteins identified by MS are stored in the ProteinRecord package.

ProteinRecord package

A new package was designed to store details of proteins identified using proteomics (Fig. 9). The class *Protein* can be referenced from MS data arising from protein identification. The protein identifier and database URL are captured in *DatabaseEntry*, and the species of origin in *OntologyEntry* (from BioOM:Description). *ProteinModification* stores information about modifications that have been observed. The type of modification, such as glycosylation or phosphorylation, is obtained from a controlled vocabulary and captured in *OntologyEntry*. The position of the modification is captured in *Location*.

IMPLEMENTATION

FGE-OM has been implemented as a relational database called RAPAD using Oracle 9i (http://www.oracle.com/ip/deploy/database/oracle9i/). The basis for a large section of the schema was RAD. RAD is used to store microarray data and is compliant with the MIAME requirements, which specify the minimum information that must be captured about a microarray experiment (Brazma *et al*., 2001). Mappings exist to convert MAGE formatted data into RAD and vice versa (MAGE-RAD Translator, MR_T). BioOM and ArrayOM are derived from MAGE-OM, and therefore tables and relationships already exist in RAD to store data modelled in these sections. We added the ProteomicsOM namespace to the database schema. First, we adapted table definitions from PEDRo to cover protein separation techniques and MS. We then designed tables to store protein spot data, along with image analysis results, using information from Gla-PSI. An interface has been created for loading data into RAPAD, developed from the RAD

Study-Annotator pages (Manduchi *et al*., 2004). 2-DE data can be viewed with a Java Applet that also acts as an access point to data from MS performed on protein spots.

We have loaded 2-DE, MS data and experimental protocols derived from a project that aims to catalogue all the expressed proteins of the protozoan parasite *Toxoplasma gondii* (Cohen *et al*., 2002). Data have also been loaded from studies to determine the changes in protein expression of host cells during a parasite invasion, compared with non-infected host cells. We can therefore demonstrate that the classes and relationships defined in FGE-OM adequately capture real data, as stored in RAPAD. RAPAD supports queries regarding the data, including

- search for all gels on which a particular protein has been identified;
- locate all the proteins that are differentially expressed in two gels, given a significance level;
- search datasets for the correlation between gene and protein expression.

In the future, we plan to create an interface enabling complex queries. We also envisage the development of data mining applications to investigate the factors that regulate global control of genes and proteins.

The RAPAD schema incorporates table definitions from RAD that capture experimental protocols. A protocol for microarrays comprises stages such as the addition of solutions, timings and mixtures applied during RNA extraction, and is stored in RAD in the table Treatment and a view (BioMaterial) on the table BioMaterialImp. The same tables are used in RAPAD to store information about the solubilization and extraction of proteins from a source of biological material in a proteomics workflow. Information about the source of material is stored in a view (BioSource) on the BioMaterialImp table. FGE-OM contains specific classes describing protein separation techniques which are defined as subclasses of *BioAssayTreatment*, and the products of separations are subclasses of *BioMaterial*. In RAPAD, a table has been created called BioAssayTreatment that is linked to individual tables storing specific information about each type of separation. The BioAssayTreatment table references the table AnalyteMeasurement, allowing a measured source of material to be specified.

Data are produced in proteomics following image analysis of 2-DE. A new table has been created, ProteomeAssay, that can be referenced by other tables storing the process of scanning, and analysis of gels (Fig. 10). A number of tables are used to store the overview of an experiment, including a parameter that is varied across a series of arrays or gels (Study, StudyDesign, StudyFactorValue and others). These tables in RAD reference the Array table and therefore have been replicated in RAPAD with an association to ProteomeAssay. MS

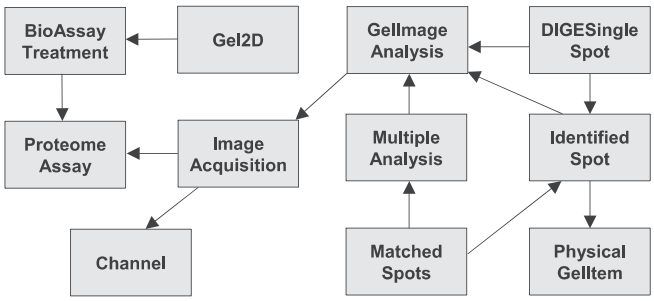

Fig. 10. Rectangles represent tables in the database storing 2-DE data. Arrows indicate a relationship between two tables, for example IdentifiedSpot has a foreign key from GelImageAnalysis.

data and protocols are stored in tables derived from the PEDRo schema, linked to gel data via BioAssayTreatment.

DISCUSSION AND CONCLUSIONS

We have developed a model called FGE-OM to represent both proteomics and microarray experiments. FGE-OM is divided into three namespaces: BioOM, ArrayOM and ProteomicsOM. We believe that BioOM can describe a generic functional genomics experiment, encompassing microarrays, 2-DE, histochemistry and others. ProteomicsOM includes classes with attributes covering 2-DE, MS and data analysis that have been integrated with BioOM, enabling additional protocols and parameters to be attached. It is our view that models describing other technologies can be added into FGE-OM without significant difficulty, allowing a unified model for functional genomics to be created. We are aware of other efforts to extend MAGE for similar purposes (Xirasagar *et al*., 2004) although taking a different approach. In light of the importance and consequence of extending standards to other areas of functional genomics, the two approaches provide valuable starting points for achieving the best result.

FGE-OM has been implemented as a relational database and tested with a set of 2-DE and MS data from studies of microbial pathogenesis. Case studies are under way to demonstrate that the schema can store proteomics and microarray data from a range of experiment types and that fast queries can be performed. The RAPAD schema is freely available to allow other developers to assess and test it in applications. RAPAD is intended for testing the capabilities of core RAD tables to store proteomics and microarray experimental protocols and data. Ultimately, separate namespaces will be created in GUS, each storing a particular technology. GUS is used to support PlasmoDB (Bahl *et al*., 2003) and ToxoDB (http://www.toxodb.org/), which are Web sites that provide access to genome and expression data from *Plasmodium falciparum* and *Toxoplasma gondii*. We will add

facilities to support protein data in GUS, allowing the expansion of ToxoDB and PlasmoDB to create a single access point to genomic, transcriptomic and proteomic data for each organism. Data derived from our studies of *Toxoplasma* will contribute to ToxoDB, with the goal of determining the entire proteome of the parasite under a range of biological conditions.

MAGE-ML is well established as a data standard for microarrays, managed by MGED, and proteomics standards are being coordinated by PSI (Orchard *et al*., 2003). FGE-OM will be supplied to PSI as a proposal for a standard format and to MGED to generate discussion about future versions of MAGE-ML. We believe that all functional genomics experiments should be annotated in a uniform manner, enabling the integration of data from a number of different experiment types. The integration will be facilitated by the convergence of microarray and proteomics models, and new formats for other types of functional genomics experiments should conform to the same standard.

ACKNOWLEDGEMENTS

We wish to thank members of the CBIL for advice on the development of FGE-OM and RAPAD. This work was supported by grants from the Medical Research Council (A.J. and E.H.), by BBSRC grant 17/513819 to J.W. and by NIH grant HG-01539 to C.J.S.

REFERENCES

Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al*. (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.

Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al*. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.

Cohen,A.M., Rumpel,K., Coombs,G.H. and Wastling,J.M. (2002) Characterisation of global protein expression by two-dimensional electrophoresis and mass spectrometry: proteomics of *Toxoplasma gondii*. *Int. J. Parasitol.*, **32**, 39–51.

Fields,S. and Song,O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

Futcher,B., Latter,G.I., Monardo,P., McLaughlin,C.S. and Garrels,J.I. (1999) A sampling of the yeast proteome. *Mol. Cell Biol.*, **19**, 7357–7368.

Jones,A., Wastling,J. and Hunt,E. (2003) Proposal for a standard representation of two-dimensional gel electrophoresis data. *Comp. Funct. Genomics*, **4**, 492–501.

Manduchi,E., Grant,G.R., He,H., Liu,J., Mailman,M.D., Pizarro,A.D., Whetzel,P.L. and Stoeckert,C.J.,Jr. (2004) RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinformatics*, **20**, 452–459.

Orchard,S., Kersey,P., Zhu,W., Montecchi-Palazzi,L., Hermjakob,H. and Apweiler,R. (2003) Progress in establishing common standards for exchanging proteomics data: the second meeting of the HUPO Proteomics Standards Initiative. *Comp. Funct. Genomics*, **4**, 203–206.

Rumbaugh,J., Jacobson,I. and Booch,G. (1999) *The Unified Modeling Language Reference Manual*. Addison-Wesley, Reading, MA, USA.

Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M. *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.*, **23**, RESEARCH0046.

Stoeckert,C.J. and Parkinson,H. (2003) The MGED ontology: a framework for describing functional genomics experiments. *Comp. Funct. Genomics*, **4**, 127–132.

Stoeckert,C., Pizarro,A., Manduchi,E., Gibson,M., Brunk,B., Crabtree,J., Schug,J., Shen-Orr,S. and Overton,G.C. (2001) A relational schema for both array-based and SAGE gene expression experiments. *Bioinformatics*, **417**, 300–308.

Taylor,C.F., Paton,N.W., Garwood,K.L., Kirby,P.D., Stead,D.A., Yin,Z., Deutsch,E.W., Selway,L., Walker,J., Riba-Garcia,I. *et al*. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.*, **21**, 247–254.

Tyers,M. and Mann,M. (2003) From genomics to proteomics. *Nature*, **422**, 193–197.

Ünlü, M., Morgan,M.E. and Minden,J.S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in cell extracts. *Electrophoresis*, **18**, 2071–2077.

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Voit,E.O. (2002) Metabolic modeling: a tool of drug discovery in the post-genomic era. *Drug Discovery Today*, **7**, 621–628.

Xirasagar,S., Gustafson,S., Merrick,B.A., Tomer,K.B., Stasiewicz,S., Chan,D.D., Yost,K.J.,III, Yates,J.R.,III, Sumner,S., Xiao,N. and Waters,M.D. (2004) CEBS Object Model for Systems Biology Data, CEBS MAGE SysBio-OM. *Bioinformatics*, Epub 25 March, 2004.